(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 May 2001 (25.05.2001)

PCT

(10) International Publication Number
WO 01/37097 A1

(51) International Patent Classification[7]: G06F 12/00, 7/36

(21) International Application Number: PCT/US00/31399

(22) International Filing Date: 15 November 2000 (15.11.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/165,621 15 November 1999 (15.11.1999) US

(71) Applicant (*for all designated States except US*): SMITHKLINE BEECHAM CORPORATION [US/US]; One Franklin Plaza, Philadelphia, PA 19103 (US).

(72) Inventor; and

(75) Inventor/Applicant (*for US only*): VICTOR, Timothy, W. [US/US]; 1020 Riverwalk Drive, Phoenixville, PA 19460 (US).

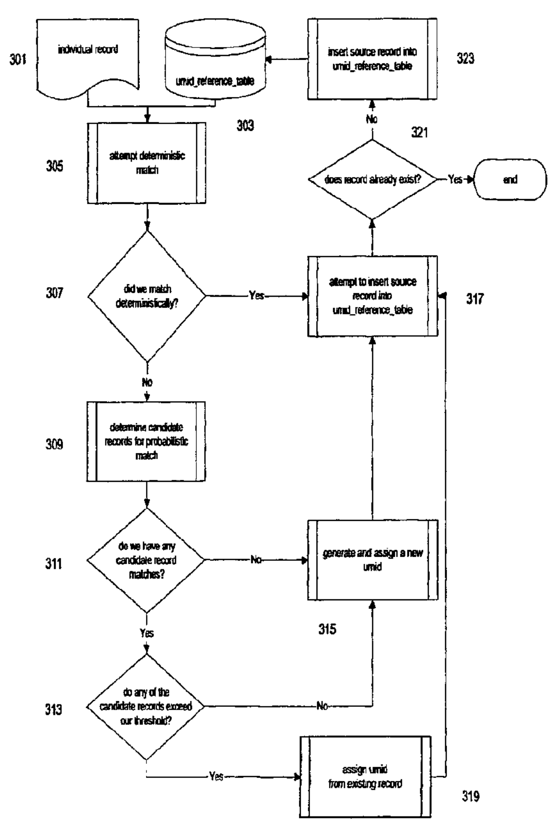
(74) Agents: KANAGY, James, M. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States (*national*): AE, AL, AU, BA, BB, BG, BR, BZ, CA, CN, CZ, DZ, EE, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KP, KR, LC, LK, LR, LT, LV, MA, MG, MK, MN, MX, MZ, NO, NZ, PL, RO, SG, SI, SK, SL, TR, TT, TZ, UA, US, UZ, VN, YU, ZA.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European

[Continued on next page]

(54) Title: METHOD FOR IDENTIFYING UNIQUE ENTITIES IN DISPARATE DATA FILES



(57) Abstract: This invention relates to a method of matching computer-based records (301) for identifying unique entities (303) both within and between disparate data files. This method of record-linkage has particular utility in the fields of epidemiology and health services research.

WO 01/37097 A1

patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*
— *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Method for Identifying Unique Entities in Disparate Data Files

Field of the Invention

This invention relates to a method of matching computer-based records for identifying unique entities both within and between disparate data files. This method of record-linkage has particular utility in the fields of epidemiology and health services research.

Background of the Invention

A custom universal identifier methodology was developed in response to the limitations of exact matching techniques. The methodology was designed to incorporate a combination of exact and probabilistic matching techniques. The term record linkage has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular entity. Integrating patient information from various sources is essential for multivariate research. The various facts concerning an individual, if brought together, form an extensive history of that individual.

There are many purposes for linking records. Examples range from obtaining more data elements about an individual by merging data from different data sources, to creating a more comprehensive name and address list by merging the names and address from several data sources. In the first case, it is important to ensure that the matching is done accurately so that the matched data truly represent a multivariate observation from a single individual. In the second, the merging is intended to ensure as complete a list as possible while eliminating duplication.

The idea of linkage records in the interest of science has a long pedigree. Fisher (Box, 1979, p. 237) lectured at a Zurich public health congress in 1929, arguing the usefulness of public records supplemented by (and presumably linked with) family data, in human genetics research. Earlier, Alexander Graham Bell exploited genealogical records and administrative records on marriages, census results and others apparently linking some sources, to sustain his familial studies of deafness (Bruce, 1973; Bell, 1906).

For many applications involving multiple databases, enough information is present to allow an accurate human judgement about whether a record from one source refers to the same case as a record from other sources. However, this is an extremely time-consuming, error-prone, and unreliable method except for small data sets. Computer methods are necessary to perform this task for a record matching exercise to be cost effective.

Summary of the Invention

The present invention is a computer-implemented system and method for creating a universal identifier for more than one record in one or more data files, the process comprising:

standardizing one or more data elements in each record;

estimating the agreement and disagreement weights employed in the probabilistic function; and

assigning a randomly generated unique identifier to each record.

In a second aspect, this invention relates to a computer-implemented system and method for concatenating records belonging to the same source within a data base or between data bases, the process comprising:

(1) creating a universal identifier for each record in one or more data files, by:

a) standardizing one or more data elements in each record;

b) estimating the agreement and disagreement weights employed in the probabilistic function; and

c) assigning a randomly generated unique identifier to each record; and

(2) concatenating records having the same unique identifier.

In yet a third aspect, this invention relates to a computer-implemented system and method for concatenating records belonging to the same source where some records have a unique identifier and new records are created, the process comprising:

(1) creating a universal identifier for each new record in one or more data files, by:

a) standardizing one or more data elements in each record;

b) estimating the agreement and disagreement weights employed in the probabilistic function; and

c) assigning a randomly generated unique identifier to each record; and

(2) concatenating records newly assigned a unique identifier with existing records having the same unique identifier.

Descripiton of the Figures

Figure 1 is a block diagram of illustrative input record components and atomic components.

Figure 2 is a flowchart of weights calculated based on chance agreement using an iterative bootstrap technique.

Future 3 is a flowchart of the process for generating randomly assigned unique identifiers.

Description of the Invention

General Overview

This invention provides a means for generating a unique identifier for records that ultimately relate back to a single source. It is particularly useful where characterizing data identifying that source expands or changes over time. Specific examples are financial data and patient data. However, in both instances, data can normally be stored in a centralised data file such as a central server only if it is adequate secured and anonymized. One way to effect this security interst is to use a trusted third party-environments. This invention has its greatest use in the trusted third-party environment.

A Trusted Third Party (TTP) service is a current way for anonymizing patient data. The data is sent to a TTP, which takes the data and replaces all patient identifiers with a new code. The TTP matches codes against the patients – it therefore knows all the codes and patients.

Working within the pervue of a TTP, or elsewhere, this invention address the step of creating and assigning a unique identifier to a record after which these records are concatenated based on the unique identifier. The creation and assignment steps have three major components: i) data standardization, ii) weight estimation, and iii) the assignment of a unique identifier, in that order.

Definitions

For the purposes of this invention, the following definitions and abbreviations are used:

μ-Probability: The probability that any random element pair will match by chance

$$\mu = \frac{n_{match}}{n_{A*B}}$$

ρ-Probability: The *reliability* of the data element. If the Element Error Rate is ≥ .99 then $\rho = 1 - EER$; Else $\rho = .99 - EER$

Agreement: A condition such that a given element pair matches exactly and both elements are known $A_{e_1} = B_{e_1}$

Agreement Weight: The weight assigned to an element pair when they agree during the record matching process

$$-\log_2\left(\frac{\rho}{\mu}\right)$$

Cartesian Product: The set of ordered pairs $A*B=\{(a,b)\mid a\in A\wedge b\in B\}$

Disagreement: A condition such that a given element pair does not exactly match and both elements are known

$$A_{e_1}\neq B_{e_1}$$

Disagreement Weight: The weight assigned to an element pair when they disagree during the record matching process.

$$\log_2\left(\frac{1-\rho}{1-\mu}\right)$$

Element Error Rate: The proportion of element pairs where at least one element is unknown, e.g., null

$$\varepsilon=\frac{n_{null}}{n_{A*B}}$$

Frequency Table: Summary of the number of times, and percentage of total different values of a variable occur

Mean: Arithmetic average

$$\overline{x_i}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

No Decision: A condition such that a given element pair where either one or both of the elements is unknown.

Random Number Assignment: Every row in the data set will be assigned a random number such that v blocks of approximately 1500 are created $\rho=\text{int}[(\upsilon*\text{P})+1]$ where ρ = Random Number, υ = Upper Bound and P = Random Function.

Threshold: The threshold utilized in probabilistic matching is a binit odds ratio with a range of $-\infty \geq x \leq \infty$.

Upper Bound: Number of strata such that the data set is divided into approximately equal rows of 1500.

$$\upsilon = \mathrm{int}\left(\frac{\text{Number of Records in Data Set}}{1500}\right)$$

As regards the computer and machine language used in this process, just about any piece of hardware capable of executing a fairly large number of calculations in shrot order will fill the bill. Any current state-of-the-art PC or server could be used. As for the operating system, UNIX is perferred, but Windows 98 or NT for Windows or the like could be used. The source code can be written in any language, though Java if preferred.

Data Standardization

The first step of this process involves the standardization of data in an input file. This standardization is required for increased precision and reliability. The input file can contain an number of variables of which one or more are or may be unique to a particular data source such as an individual. Examples of useful variables are: member identifier, drivers' license number, social security number, insurance company code number, name, gender, date of birth, street address, city, state, postal code, citizenship. In addition, some identifiers can be further distilled down into their basic, or atomic, components. Figure 1 illustrates the use of selected input record components and atomic components of some records that are amenable to such further distillation. Referring to Figure 1, Input Record 100 illustrates data which can be used as the basis for assigning a unique identifier, and how that data can be broken out inot its atomic and subatomic components exemplified by Street Address 110, Date of Birth 120 and Name 130.

During the standardization process, all character data is preferably transformed to a single case. For example they are transformed to uppercase. So for instance, first names are standardized to uppercase, e.g., {BOB, ROB, ROBBY} = ROBERT. Common names for cities and streets may be transformed to the postal code, e.g., in the U.S. to United States Postal Service standard. In the latter instance this can be done using industry standard CASS certified software.

Weight Estimation

A fundamental component of this algorithm is the process of estimating the agreement and disagreement weights necessary for the probabilistic function. Weights are calculated based in probabilities of chance agreement using an iterative bootstrap technique. Figure 2 provides a flow of the process.

The first step in the weight estimation process is to determine the number of strata required such that the data set can be divided into approximately equal blocks of 1500 rows (Fig. 2 - 201-219), see equation 1.

$$\upsilon = \mathrm{int}\left(\frac{\text{Number of Records in Data Set}}{1500}\right) \quad (1)$$

The source file is then scanned and the records are assigned a random number between 1 and υ. A data matrix is created containing a Cartesian product of records with a random number of 1 assigned. The resulting matrix is then scanned. Each element pair within each record pair is assessed and assigned a value in the following manner:

$$e_n = \begin{cases} 1 \text{ if } A_{e_n} = B_{e_n} \text{ (Agreement)} \\ 0 \text{ if } A_{e_n} = \mathit{Null} \text{ and/or } B_{e_n} = \mathit{Null} \text{ (No decision)} \\ -1 \text{ if } A_{e_n} \neq B_{e_n} \text{ (Disagreement)} \end{cases} \quad (2)$$

where A_{e_n} is the nth element from record A

Once the matrix has been fully assessed, percentages for each e_n are tabulated and stored. This process is repeated for 15 iterations.

Mean percentages of *Agreements* and *No Decisions* are calculated for each data element (Fig. 2 - 221). The ρ probability, or the reliability, for each data element is then calculated, see equation 3.

$$\mathit{let}\ \varepsilon = \overline{x}_{\mathit{Percent\ No\ Decision}}$$

$$\rho = \begin{cases} \mathit{if}\ \varepsilon \geq .99\ \mathit{then}\ 1 - \varepsilon \\ \mathit{else}\ .99 - \varepsilon \end{cases} \quad (3)$$

The μ probability, or the probability that element *n* for any given record pair will match by chance, is calculated (Fig. 2 - 223), see equation 4.

$$\mu = \overline{x}_{\mathit{Percent\ Agreement}} \quad (4)$$

From the ρ and μ probabilities, the disagreement and agreement weight formula are calculated (Fig. 2 - 225)employing equations 5 and 6 respectively.

$$Disagreement = \log_2\left(\frac{1-\rho}{1-\mu}\right) \quad (5)$$

$$Agreement = \log_2\left(\frac{\rho}{\mu}\right) \quad (6)$$

Unique Identifier Assignment

The final stage of this process is the action of uniquely identifying entities within the input data set. Figure 3 provides an overview of this process.

Each record from the input file is evaluated against a reference file to determine if the entity represented by the data has been previously identified using a combination of deterministic and probabilistic matching techniques. If it is judged that the entity is already represented in the reference set, the input record is assigned the unique identifier (UID) from the reference record that it has matched against. If it is judged that the entity represented by data is not yet in the reference set, a new UID is randomly generated and assigned. Random numbers are generated in whatever language the process is being implemented.

After the UID assignment occurs, the input record is evaluated, in it's entirety, to determine if the record is a unique representation of the entity not already contained in the reference table. If it is a new record, then it is inserted into the reference table for future use.

Deterministic Matching Technique

The deterministic matching technique employs simple Boolean logic. Two records are judged to match if certain criteria are met, such as the following:

First Name Matches Exactly

Last Name Matches Exactly

Date of Birth Matches Exactly

Social Security Number OR Member Identifier Matches Exactly

If two records satisfy the criteria for deterministic matching, no probabilistic processing occurs. However, if no deterministic match occurs, the input record is presented for a probabilistic match.

Probabilistic Matching Technique

The first step in the probabilistic matching process is to build a set of candidate records from the reference table based on characteristics of specific elements of the input record. This process is referred to as blocking, the set of candidate records is referred to as the blocking table. All data sets do not use the same characteristics, the elements used in this process are determined through data analysis. However, it is suggested that blocking

variable consist of those elements that are somewhat unique to an element, e.g., social security number, or a combination of date of birth and last name.

Upon completion of the construction of the blocking table, each element for each candidate record is compared against its corresponding element from the input record. See equation 7 for the scoring mechanism.

$$w_n = \begin{cases} \text{Agreement Weight if } A_{e_n} = B_{e_n} \\ 0 \text{ if } A_{e_n} = \text{Null and/or } B_{e_n} = \text{Null} \\ \text{Disagreement Weight if } A_{e_n} \neq B_{e_n} \end{cases} \quad (7)$$

where A_{e_n} is the nth element from record A

A composite weight is then calculated for all candidate records, see equation 8.

$$W = \sum_{i=1}^{N} w_i \quad (8)$$

The candidate record with the highest composite weight is then evaluated against a predefined threshold. If the weight meets or exceeds the threshold, the candidate record is judged to match the input record. If the weight does not exceed the threshold, it is assumed that the input record represents an entity not yet included in the reference set.

What is claimed is:

1. A computer-implemented system and method for creating a universal identifier for more than one record in one or more data files, the process comprising:

standardizing one or more data elements in each record;

estimating the agreement and disagreement weights employed in the probabilistic function; and

assigning a randomly generated unique identifier to each record.

2. A computer-implemented system and method for concatenating records belonging to the same source within a data base or between data bases, the process comprising:

(A) creating a universal identifier for each record in one or more data files, by:

a) standardizing one or more data elements in each record;

b) estimating the agreement and disagreement weights employed in the probabilistic function; and

c) assigning a randomly generated unique identifier to each record; and

(B) concatenating records having the same unique identifier.

3. A computer-implemented system and method for concatenating records belonging to the same source where some records have a unique identifier and new records are created, the process comprising:

(A) creating a universal identifier for each new record in one or more data files, by:

a) standardizing one or more data elements in each record;

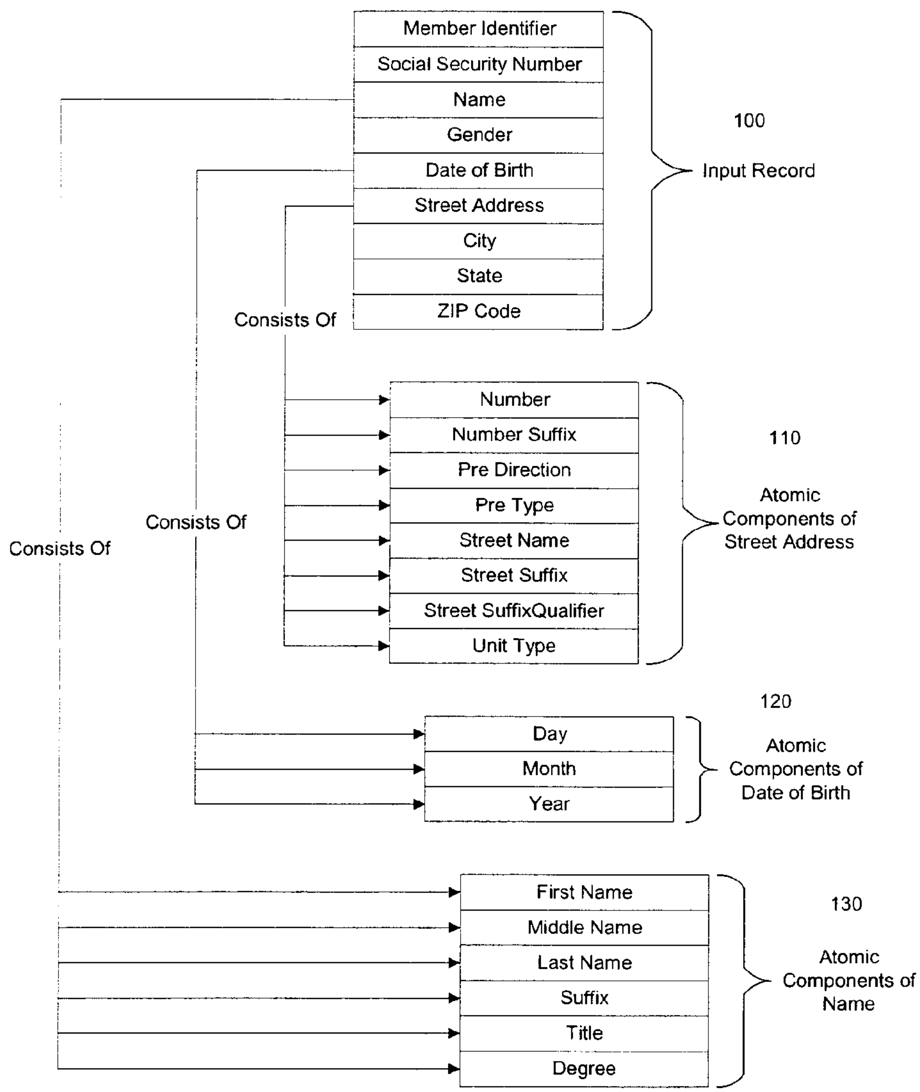
b) estimating the agreement and disagreement weights employed in the probabilistic function; and

c) assigning a randomly generated unique identifier to each record; and

(B) concatenating records newly assigned a unique identifier with existing records having the same unique identifier.

4. A method for assigning a unique identification number to a source or owner data as described herein.

Figure 1



Member Identifier
Social Security Number
Name
Gender
Date of Birth
Street Address
City
State
ZIP Code
100
Input Record
Consists Of
Number
Number Suffix
Pre Direction
Pre Type
Street Name
Street Suffix
Street SuffixQualifier
Unit Type
110
Atomic Components of Street Address
Consists Of
Consists Of
Day
Month
Year
120
Atomic Components of Date of Birth
First Name
Middle Name
Last Name
Suffix
Title
Degree
130
Atomic Components of Name

Figure 2

2/3

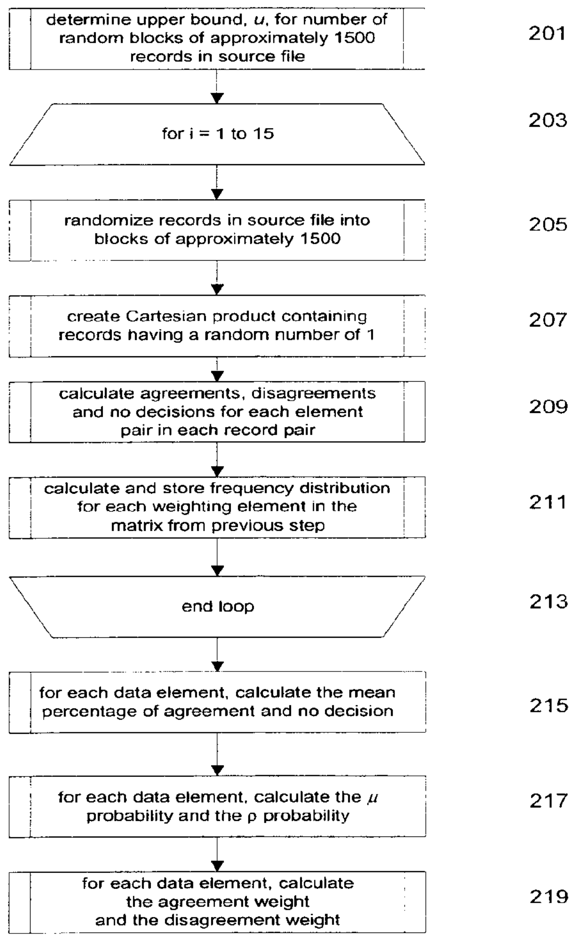
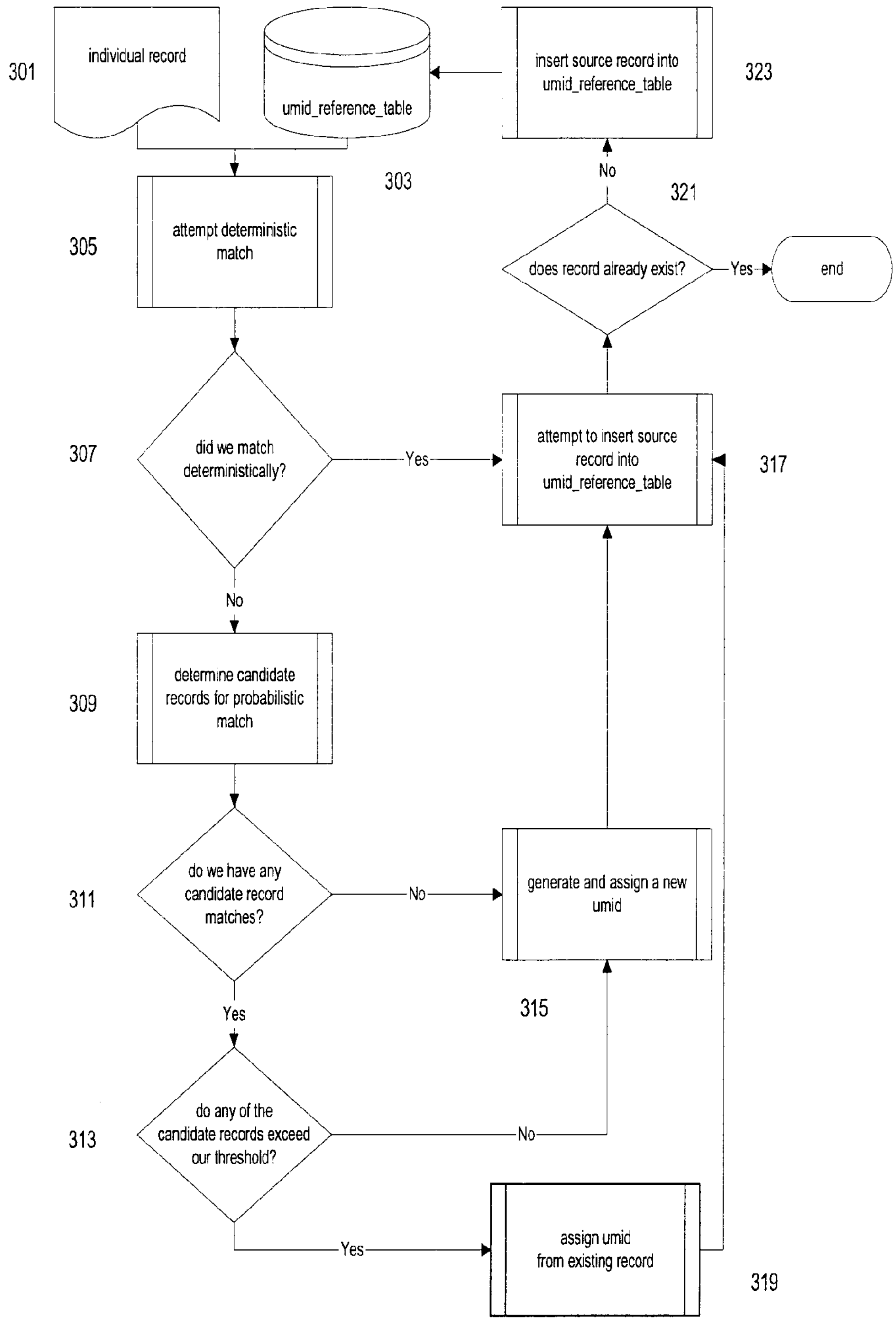

determine upper bound, u, for number of random blocks of approximately 1500 records in source file
201
for i = 1 to 15
203
randomize records in source file into blocks of approximately 1500
205
create Cartesian product containing records having a random number of 1
207
calculate agreements, disagreements and no decisions for each element pair in each record pair
209
calculate and store frequency distribution for each weighting element in the matrix from previous step
211
end loop
213
for each data element, calculate the mean percentage of agreement and no decision
215
for each data element, calculate the μ probability and the ρ probability
217
for each data element, calculate the agreement weight and the disagreement weight
219

Figure 3

3/3


301
individual record
umid_reference_table
303
insert source record into umid_reference_table
323
No
321
305
attempt deterministic match
does record already exist?
Yes
end
307
did we match deterministically?
Yes
attempt to insert source record into umid_reference_table
317
No
309
determine candidate records for probabilistic match
311
do we have any candidate record matches?
No
generate and assign a new umid
Yes
315
313
do any of the candidate records exceed our threshold?
No
Yes
assign umid from existing record
319

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/31399

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 12/00; G06F 7/36
US CL : 707/2, 3, 6, 104; 710/7
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/2, 3, 6, 104; 710/7

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EAST, IEL, ACM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 4,821,184 A (CLANCY et al) 11 April 1989	1-3
A	US 5,594,889 A (COLGATE et al) 14 January 1997	1-3
A	US 5,487,164 A (KIRCHHOFER et al) 23 January 1996	1-3
A	US 5,668,897 A (STOLFO) 16 September 1997	1-3

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 DECEMBER 2000	16 MAR 2001

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	UYEN LE James R. Matthews
Facsimile No. (703) 305-3230	Telephone No. (703) 305-3900

Form PCT/ISA/210 (second sheet) (July 1998)★

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/31399

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 4
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

Claim 4 recites "a method for assigning a unique identification number to a source or owner data as described herein". However, there is no description of the claimed method. Therefore, the limitation can not be ascertained.

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.